Selective pressure against horizontally acquired prokaryotic genes as a driving force of plastid evolution

Briardo **Llorente**[1,2,*], Flavio S. J. **de Souza**[2,+], Gabriela **Soto**[2,+], Cristian **Meyer**[2], Guillermo D. **Alonso**[2,3], Mirtha M. **Flawiá**[2,3], Fernando **Bravo-Almonacid**[2,4], Nicolás D. **Ayub**[5,6] and Manuel **Rodríguez-Concepción**[1,*]

[1] Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, 08193 Barcelona, Spain.

[2] Instituto de Investigaciones en Ingeniería Genética y Biología Molecular Dr. Héctor Torres, Consejo Nacional de Investigaciones Científicas y Técnicas (INGEBI-CONICET), C1428ADN Buenos Aires, Argentina.

[3] Departamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, C1428EGA Buenos Aires, Argentina.

[4] Departamento de Ciencia y Tecnología. Universidad Nacional de Quilmes, B1876BXD Bernal, Argentina.

[5] Instituto de Genética Ewald A. Favret, Centro de Investigación en Ciencias Veterinarias y Agronómicas, Instituto Nacional de Tecnología Agropecuaria (CICVyA-INTA), B1712WAA Castelar, Argentina.

[6] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), C1033AAJ Buenos Aires, Argentina.

[+] These authors contributed equally to this work.

* Corresponding authors:

BL, briardo.llorente@cragenomica.es

MRC, manuel.rodriguez@cragenomica.es

The plastid organelle comprises a high proportion of nucleus-encoded proteins that were acquired from different prokaryotic donors via independent horizontal gene transfers following its primary endosymbiotic origin. What forces drove the targeting of these alien proteins to the plastid remains an unresolved evolutionary question. To better understand this process we screened for suitable candidate proteins to recapitulate their prokaryote-to-eukaryote transition. Here we identify the ancient horizontal transfer of a bacterial polyphenol oxidase (PPO) gene to the nuclear genome of an early land plant ancestor and infer the possible mechanism behind the plastidial localization of the encoded enzyme. Arabidopsis plants expressing PPO versions either lacking or harbouring a plastid-targeting signal allowed examining fitness consequences associated with its subcellular localization. Markedly, a deleterious effect on plant growth was highly correlated with PPO activity only when producing the non-targeted enzyme, suggesting that selection favoured the fixation of plastid-targeted protein versions. Our results reveal a possible evolutionary mechanism of how selection against heterologous genes encoding cytosolic proteins contributed in incrementing plastid proteome complexity from non-endosymbiotic gene sources, a process that may also impact mitochondrial evolution.

Introduction

A large number of plant genes encoding plastid proteins originated via independent horizontal gene transfer (HGT) events from prokaryotes other than *Cyanobacteria*[1,2], the accepted endosymbiotic precursor of modern-day plastids[3]. The vast majority of those genes are nuclear, and therefore their protein products are translated in the cytosol and subsequently imported into the plastid. This in itself is remarkable, because prokaryotic genes lack the information needed to target their encoded proteins to eukaryotic organelles and hence must have acquired it following insertion in the nuclear genome and transcriptional activation. Despite the profound evolutionary importance of this process, we know little about the forces underlying the targeting of these alien proteins to the plastid.

We addressed this conundrum using a novel approach that is based on the idea that studying the ancestral (cytosolic) and modern (plastidial) subcellular localization of HGT-derived proteins *in vivo* may provide clues about the evolutionary mechanisms that ultimately resulted in their organellar targeting. We accordingly screened for candidate proteins of non-photosynthetic function that were present exclusively in prokaryotes and plants, and whose plastidial localization in the later could not be explained by a strict functional requirement. A uniquely suited candidate for a case study was found to be polyphenol oxidase (PPO; Enzyme Commission 1.14.18.1 or 1.10.3.1), a bacterial enzyme[4] whose nucleus-encoded plant homologues are almost ubiquitously present in plastids[5]. Although the physiological functions of PPO are not fully understood in plants, its capability to oxidize phenolic compounds (e.g. phenylpropanoids) into highly reactive quinones is thought to function as an "armed bomb" defence mechanism against phytophagous organisms functioning only when cell damage occurs and plastids break[6-8]. In bacteria, PPO enzymes are involved in the production of melanin-like polymers that protect bacterial cells and spores against UV radiation, heat, metal ions, oxidants, enzymatic hydrolysis, antimicrobial compounds and phagocytosis[9].

We present here evidence indicating that an early common ancestor of land plants acquired a bacterial gene coding for PPO through an ancient HGT event during the plant colonization of land. Recapitulating the bacteria-to-plants transition of PPO in the model plant *Arabidopsis thaliana* (Brassicaceae; hereafter Arabidopsis) suggests that the plastidial targeting of PPO in plants is likely a consequence of selective pressures acting against the deleterious effects caused by a PPO enzyme functioning in the cytosol. Given that biological processes in eukaryotic cells are highly compartmentalized, selection against horizontally acquired genes encoding non-targeted heterologous

proteins may represent a general route by which novel organellar protein versions can arise.

Results

Plants horizontally acquired a bacterial *PPO* gene during colonization of land.

Genes coding for PPO have been found in bacteria, plants and fungi[10,11], but their phylogenetic relations have not been studied in detail. At the protein level, PPOs of all three groups harbour a catalytic tyrosinase (TYR) domain (PFAM00264) towards the N-terminus, found not only in PPOs but also in several proteins of all branches of life[12] (**Fig. 1a**). In plants, the C-terminal half of PPO has distinctive DWL (PFAM12142) and KFDV (PFAM12143) domains[5]. Both DWL and KFDV domains are identified as such by the Conserved Domains Database (CDD)[13] in PPOs of various unrelated bacteria like the free-living betaproteobacterium *Chitinomonas koreensis*, the plant pathogen betaproteobacterium *Ralstonia solanacearum*, the gammaproteobacterium *Rheinheimera* sp. and the cyanobacterium *Gloeocapsa* sp. Some bacterial PPOs possess only one of these domains, while PPOs of many other bacteria as well as all fungal PPOs lack recognizable DWL and FKDV domains (**Fig. 1a** and **1b**). Sequence alignments of PPO peptides from plants, bacteria and fungi show that the TYR domain can be readily aligned. The DWL and the KFDV domains, although only distantly related, can also be aligned among these proteins, indicating that the PPOs of these organisms are phylogenetically related (**Supplementary Fig. S1**). Notably, sequences encoding proteins with a PPO-like structure could not be retrieved from the complete genomes of the green algae *Volvox carteri*, *Chlamydomonas reinhardtii*, *Micromonas putida* and *Chlorella* sp., as well as from the genomes of red algae and other distant organisms like stramenopiles or haptophyte algae. Although some representatives of these groups have proteins containing TYR domains, DWL and KFDV domains could not be identified in any of them (**Fig. 1a**).

The survey described above indicates that PPO proteins are only found in land plants, fungi and some bacteria. To analyse the phylogenetic relations between these proteins, we focused on the TYR domain. A Maximum Likelihood analysis of the relationships between PPO tyrosinase (PPO-TYR) domains of land plants, bacteria and fungi reveals that land plant PPO-TYR domains form a well-supported monophyletic group with the PPO-TYR of several bacteria (**Fig. 1b**). Protein structure predictions of *R. solanacearum* (WP_003278615), *P. patens* (AAX69084) and *V. carteri*

(XP_002956492) further illustrate a greater structural similarity among land plant and bacterial TYR domains (**Fig. 1c**). The PPO-TYR domains of fungi are grouped in a separate, well-supported branch (**Fig. 1b**). Although green algae lack bona fide PPO proteins, we included in our phylogenetic analysis the only TYR domains that could be retrieved from the proteomes of *Chlamydomonas* and *Volvox*. As shown in **Fig. 1b**, the green algae TYR domains do not cluster with land plants, suggesting that land plants have not inherited PPO-TYR domains from a green algal ancestor by vertical inheritance. The patchy phylogenetic distribution of *PPO* genes as well as the phylogenetic incongruences observed for the PPO-TYR domains support the conclusion that PPO proteins are of ancient origin but have been subjected to many different HGT events during evolution. In particular, the fact that PPO-TYR domains of land plants, including the basal groups Bryophyta (*Physcomitrella patens*) and Lycophyta (*Selaginella moellendorffii*), are all grouped together with bacteria and that *PPO* genes cannot be found in glaucophyte, red and green algal genomes, strongly suggest that a *PPO*-like gene was transferred from a bacterium to an early land plant ancestor circa 450-500 million years ago during the transition of plants to live on land[14] (**Fig. 1d**). Additional evidence supporting the *PPO* bacteria-to-plants HGT is the observation that, compared to the average of plant genes, which typically contain 5-6 introns[15,16], plant *PPOs* have substantially fewer or no introns (0-2 introns)[5], a characteristic feature of prokaryotic horizontally acquired genes[17].

Recapitulation of the *PPO* prokaryote-to-eukaryote transition.

Having defined the bacteria-to-plants HGT of *PPO*, we set to experimentally recreate its horizontal acquisition by land plants. Our aim was to generate a model to study the effect of incorporating a nuclear gene coding for a protein of prokaryotic origin (PPO) that initially lacked information for a defined subcellular localization and eventually evolved a plastidial destination. Given that inferred ancestral proteins and proteins having evolved in distant lineages may result inactive or cause a stress derived from maladapted cell components when introduced into a distantly related host[18,19], we decided to use a modified plant PPO to simulate the presumed ancestral cytosolic state of the enzyme rather than a resurrected or a bacterial protein. To ensure that the enzyme used would be active, we chose a *PPO* gene from *Solanum tuberosum* that we previously confirmed to encode an active PPO enzyme[20]. The plastid-targeting signal sequence of the *PPO* gene was identified *in silico* using ChloroP[21]. Then, two versions of this gene were cloned: one lacking the predicted plastid-targeting

signal (version PPO_A, for enzyme with Ancestral localization) and one retaining it (version PPO_M, for enzyme with Modern localization) (**Fig. 2a**). The recombinant proteins were expressed in *Escherichia coli* and the activity of PPO_A and PPO_M was confirmed *in vivo* based on the ability of both proteins to induce melanin synthesis (**Fig. 2b**). To determine the subcellular localization of PPO_A and PPO_M in plant cells, a GFP (green fluorescent protein) tag was fused to the C-terminal end of both PPO versions and the corresponding constructs were transiently expressed in *Nicotiana benthamiana* leaves. Laser-scanning confocal microscopy analyses showed that PPO_A presents a cytosolic distribution pattern while PPO_M is detected only in plastids (chloroplasts) (**Fig. 2c**), as expected.

To experimentally study the PPO bacteria-to-plants transition (albeit in a modern context), we generated stably transformed Arabidopsis plants by *Agrobacterium*-mediated transformation with constructs for constitutive PPO_A or PPO_M expression. In agreement with the conclusion that PPO activity is not absolutely required for photosynthesis or any other primary plant function[5,6,20,22], Arabidopsis does not contain any *PPO* genes[5] and thus it has lived without PPO for several million years[23]. We therefore chose Arabidopsis as an optimal plant system to have an approximation of the physiological impact of acquiring a *PPO* gene *ab initio*. After isolating independent homozygous lines containing a single T-DNA insertion in the nuclear genome, we selected three PPO_A and three PPO_M lines with different *PPO* gene-expression (**Fig. 3a**) and activity (**Fig. 3b**) levels. Expressing the cytosolic PPO_A led to reduced growth rates, while expressing the plastid-localized PPO_M had no evident phenotypic effect on transgenic lines compared to untransformed controls (**Fig. 3c-e**). The expression and activity of the cytosolic version inversely correlated ($r = -0.97$ and $r = -0.99$, respectively) with the surface area of rosette leaves (**Fig. 3d**), supporting the conclusion that the observed deleterious phenotype in PPO_A lines was directly caused by the presence of such an extraplastidial PPO activity. Plants transformed with PPO_A were also considerably delayed in terms of bolting time (**Fig. 3e**), which again correlated with its expression and activity ($r = 0.93$ and $r = 0.99$, respectively).

At the metabolic level, the qualitative profile (i.e. number of HPLC peaks) of phenylpropanoid compounds such as flavonols (**Fig. 4a**) and anthocyanins (**Fig. 4b**) was found to be very similar in PPO_A, PPO_M, and untransformed plants, However, the presence of an active cytosolic PPO enzyme in PPO_A lines had a strong quantitative impact on anthocyanin levels, which were doubled relative to PPO_M plants or wild-type controls (**Fig. 4c**), Anthocyanins are purple pigments widely used as visual

markers of plant cell stress. For example, treatment with norflurazon generates albino seedlings in which the absence of photosynthetic pigments (carotenoids and chlorophylls) causes a strong photooxidative stress and a visually detected accumulation of anthocyanins that again is higher in PPO_A lines (**Fig. 4d**). Together, these results suggest that PPO enzymes can act on unidentified cytosolic substrates to eventually unchain a stress response. If such response is sustained (e.g. in lines constitutively expressing the extraplastidial PPO_A enzyme), a negative impact on plant growth occurs.

Discussion

Despite the relevant contribution of HGT-acquired prokaryotic proteins to the establishment of the plastid proteome[1,2], the evolutionary forces that promoted their targeting to the plastid have remained an unresolved evolutionary question. Focusing on a carefully selected candidate gene (*PPO*) for a case study, we have aimed to shed some light on this subject.

Our results strongly suggest that an early common ancestor of land plants horizontally acquired a gene coding for PPO of bacterial origin, possibly as a consequence of life in close contact with soil bacteria (**Fig. 1d**). An early eukaryotic origin of *PPO* is less parsimonious and would require, at least, multiple independent *PPO* losses in the glaucophyte, red algae and green algae lineages, as well as in the animal lineage. Furthermore, since the last common ancestor of the supergroup Plantae[3] originated more than 1.2 billion years ago[24], this scenario would also imply that the *PPO* gene was maintained for more than 700 million years only in the algal lineage leading to land plants to be then lost after the colonization of land. Interestingly, the land plant acquisition of *PPO* corresponds to the period when the phenylpropanoid biosynthetic pathway (a rich source of phenolic metabolites) emerged in land plants, also in part by the gain of horizontally acquired genes from soil bacteria or fungi[25]. These coincidental events most likely allowed unprecedented interactions between PPO and the phenylpropanoid pathway, setting the ground for evolutionary innovations that led to the development of defensive and secondary metabolic functions for PPO in land plants. For instance, keeping the PPO enzyme separated from phenylpropanoid substrates that could be converted into toxic products for invaders could have represented a novel resource-efficient defence mechanism that only becomes active when cells are attacked and components of different subcelular compartments are mixed.

It has been established that translocation of genes from the essentially prokaryotic plastid genome to the nucleus involves post-insertional genomic rearrangements rather than random integration into a genomic location conferring instant functionality[26-29]. On the other hand, the acquisition of plastid targeting peptides is a process that can evolve with relative ease from existing genome sequences[30,31]. According to these data, it is presumed that ancestral prokaryote-to-eukaryote horizontally transferred sequences typically involved an initial integration into the nuclear genome followed by the incorporation of eukaryotic transcriptional regulatory elements and, in many cases, the eventual acquisition of subcellular localization signals. Therefore, for most HGT events resulting in plastid-localized proteins, there must have been an evolutionary time when populations contained both the ancestral (i.e. non-targeted, cytosolic) version of the protein and the modern (i.e. targeted, plastid-localized) version that was ultimately retained. From a Darwinian fitness viewpoint, and independently of the defence mechanism proposed for PPO in land plants, our results allow proposing that the ancestral bacterial gene coding for a non-targeted PPO integrated in the nuclear genome of an early common ancestor of land plants, became transcriptionally active, and subsequently evolved a plastid-targeting sequence, resulting in a more beneficial variant that was hence selected. Similar to that observed in bacteria[9], it is possible that the acquisition of a non-targeted (i.e. cytosolic) PPO by early land plants might have provided the ability to synthesize phenolic polymers or other compounds with protective effects. This putative adaptive advantage might have been turned into a disadvantage once the transition from aquatic to terrestrial environments was completed. Alternative evolutionary trajectories that also fit our results would be that, when acquired by the land-plant ancestor, the bacterial *PPO* gene integrated by chance in a genomic location simultaneously conferring both transcriptional activation and plastid targeting or that the acquisition of a plastid localization signal preceded the transcriptional activation of the gene. Under these hypotheses, eventual mutations or genomic rearrangements resulting in the loss of the plastid-localization signal would lead to a disadvantageous phenotype and the cytosolic PPO would not be retained. In either case, the finding that PPO_A triggers a stress response (as deduced from the accumulation of anthocyanins) that could eventually lead to reduced growth rates when compared with PPO_M implies that directional selection would increase the frequency of PPO_M relative to PPO_A after the horizontal acquisition of the bacterial gene, ultimately favouring the fixation of the plastidial enzyme within the population.

In stark contrast to the common assumption that newly acquired genes conferring a strongly negative fitness effect will irreversibly undergo gene degeneration or loss, we propose that the fortuitous acquisition of organelle targeting sequences can provide an alternative fate to such genes. The argument is made that a horizontally acquired prokaryotic gene coding for a delocalized protein could have a deleterious effect that is only context-specific in a eukaryotic cell where different biological processes are highly compartmentalized. The targeting of the foreign protein to an organelle such us the plastid could alleviate the deleterious effect, hence allowing the gene to endure in the recipient genome and thereafter serve as raw material for the evolution of novel functions. The plausible generality of the proposed evolutionary mechanism is supported by the well-known fact that cytosolic localization of heterologous proteins is often deleterious to plant growth when compared with plastid-localized versions[32-35], probably due to interference with cytosolic metabolic pathways and the enormous capacity of the plastid to accumulate heterologous proteins.

A fundamental goal in biology is to elucidate the forces leading to evolutionary change. Previous work has revealed the unexpected broad contribution of prokaryotic proteins acquired via independent HGT events to the plastid proteome[1,2], the processes leading to gene functionalization of prokaryote-like (plastid) sequences in the nucleus[26,27] and the evolutionary acquisition of plastid targeting sequences[30,31]. By studying the likely evolutionary history followed by a horizontally transferred bacterial gene after acquisition by land plants hundreds of millions of years ago, we have experimentally shown that selective pressure against heterologous genes encoding cytosolic proteins may be an evolutionary force driving plastid proteome complexity from non-endosymbiotic gene sources. Finally, it has not escaped our notice that, as with the plastid, the inference we have postulated may also apply to certain horizontally transferred prokaryotic genes whose encoded proteins have been ultimately targeted to the mitochondrion[36-38].

Methods

Gene sampling

Extensive BLAST[39] searches spanning the diversity of eukaryotic and prokaryotic life were performed to identify putative eukaryotic and prokaryotic PPO homologues. Additional searches were also performed excluding land plants (embryophytes, taxid: 3193). The searches for amino acid sequences

were conducted against the GenBank non-redundant protein sequences database using the BLASTp tool of the NCBI (National Center for Biotechnology Information) BLAST server. The PPO-TYR domain was also used to search the genomes of unicellular eukaryotes deposited in the Joint Genome Institute (http://genome.jgi-psf.org/) and Phytozome (www.phytozome.net). To check the robustness of the gene sampling additional searches using PSI-BLAST[39] (5 iterations), tBLASTn and HMMER[40] were performed.

Phylogenetic analysis

Tyrosinase domains of PPO proteins and other TYR-containing proteins were aligned with ClustalW[41] and edited with Jalview 2.8[42] to minimize gaps. Information about sequences used is described in detail in **Supplementary Table S1**. The tree in **Fig. 1b** used the LG protein substitution model[43], but JTT[44] and Blosum62[45] yielded essentially the same results. Bootstrapped maximum likelihood phylogenetic trees were constructed with Phyml 3.0[46,47]. Gamma correction was used to account for heterogeneity in evolutionary rates with four discrete classes of sites, an estimated alpha parameter and an estimated proportion of invariable sites as implemented in Phyml. A total of 100 bootstrap replicates were used to assess robustness, and tree topology was optimized by subtree pruning and regrafting (SPR). The tree was visualized and edited with TreeDyn[48]. The **Fig. 1d** scheme summarizing and simplifying our current understanding of the supergroup Plantae evolution was inspired by those recently published[3,49,50].

Protein domain architecture comparison

Genes coding for proteins with tyrosinase domains were compared by examining their corresponding protein domain architectures using CDART (Conserved Domain Architecture Retrieval Tool: http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi)[51], and the CDD (Conserved Domains Database: http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml)[13] of the NCBI.

Structural modelling

The structural modelling was performed using the intensive modelling mode of the Protein Homology/analogy Recognition Engine (Phyre) Version 2.0 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index)[52].

Plasmids, plant material and growth conditions

Full-length cDNAs encoding *Solanum tuberosum PPO* (GenBank accession number: U22921) or a *PPO* version lacking localization signal were amplified from cDNA and cloned into plasmid pDONR207 using Gateway technology (Invitrogen, California, USA). The *in silico* detection of the subcellular localization signal was identified using ChloroP[21] (http://www.cbs.dtu.dk/services/ChloroP/). Sequences were then subcloned into plasmids pGWB405[53], pET28[54] (modified for Gateway-compatible cloning) and pB7FWG2[55]. Constructs were confirmed by restriction mapping and DNA sequence analysis. Information about primers used is described in detail in **Supplementary Table S2**. Constructs in pGWB405 were used for transient expression in *Nicotiana benthamiana* leaves[56]. Constructs in pET28 were used for protein expression in *Escherichia coli*. Constructs in pB7FWG2 were used for stably *Agrobacterium tumefaciens*-mediated transformation of *Arabidopsis thaliana* plants (ecotype Col-0)[57]. Homozygous Arabidopsis lines containing a single T-DNA insertion were selected based on the segregation of the phosphinotricin (PPT) resistance marker. Initial segregation analysis was performed by seeding plants on sterile Murashige and Skoog (MS) medium plates supplemented with 20 µg/mL PPT (Duchefa, Haarlem, Netherlands). A total of 25 PPO_M and 29 PPO_A homozygous lines were isolated; from which three independent lines of each PPO version were selected for further characterization. Arabidopsis plants were grown on soil in a climate controlled growth chamber (22ºC, 65-70% RH, and 60 µmol m^{-2} s^{-1} photosynthetically active radiation) under 8 h of light/16 h of dark photoperiod for 4 weeks. For phenylpropanoid profiling, Arabidopsis seeds were surface-sterilized and sown on Petri plates with sterile MS medium containing 1% agar and 30 g/L sucrose. When indicated, plates were supplemented with 5 µM norflurazon. After stratification for 3 days at 4ºC in the dark, the plates were incubated in growth chambers for 20 days at 22°C under continuous light (130 µmol m^{-2} s^{-1} photosynthetically active radiation).

Expression of recombinant PPO_A and PPO_M in *E. coli* cells

Following transformation of competent *E. coli* cells strain Rosetta 2 (DE3) (Novagen, Merck KGaA, Darmstadt, Germany) with purified plasmids pET28-PPO_A and pET28-PPO_M, the activity of PPO_A and PPO_M was confirmed by growing the transformed *E. coli* cells for three days at 28ºC on LB agar plates supplemented with 34 µg/mL chloramphenicol, 25 µg/mL kanamycin, 0.5 mM IPTG, 40 µg/ml $CuSO_4$,

and 600 µg/ml chlorogenic acid, similarly to methods previously described[58] to detect melanin formation. PPO_A and PPO_M colonies were dark because the cells were able to produce melanin-like polymers.

Confocal laser-scanning microscopy.

Subcellular localization of GFP fluorescence and chlorophyll fluorescence was determined with an Olympus FV 1000 confocal laser-scanning microscope (Olympus, Tokyo, Japan) using an argon laser for excitation (at 488 nm) and a 500–510 nm filter for detection of GFP fluorescence and a 610–700 nm filter for detection of chlorophyll fluorescence.

Measurement of rosette leaves surface area

The surface area of rosette leaves was measured using Photoshop (Adobe, California, USA) and GraphPad Prism 5.0a (GraphPad Software, California, USA) by analysing digital images (acquired from above) of 4-week-old plants and comparing them with size standard series ranging from 4 to 16 cm^2.

Transcripts analysis

Leaf samples from 4-week-old plants were harvested and total RNA was extracted using the Maxwell 16 LEV simplyRNA Tissue Kit (Promega, Wisconsin, USA) according to the manufacturer's instructions. Purified RNAs were quantified by spectroscopy using a NanoDrop apparatus (Thermo Scientific, Massachusetts, USA) and RNA integrity was evaluated by agarose gel electrophoresis. The First Strand cDNA Synthesis Kit (Roche, Basel, Switzerland) was used to generate cDNA according to the manufacturer's instructions, using 50 pmol of an anchored poly(dT) primer [$d(T_{18}V)$] and 2 µg of total RNA. The relative mRNA abundance was evaluated via quantitative reverse transcription PCR (RT-qPCR) in a total reaction volume of 20 µl using LightCycler 480 SYBR Green I Master (Roche, Basel, Switzerland) on a LightCycler 480 Real-Time PCR System (Roche, Basel, Switzerland) with 0.3 µM of each specific sense and anti-sense primers. Three independent biological replicates of each sample and three technical replicates of each biological replicate were performed and the mean values were considered for further calculations. The normalized expression of *PPO* was calculated as described[59] using *UBC* (Arabidopsis ubiquitin-conjugating enzyme gene At5g25760, GenBank

accession number: DQ027035) as the endogenous reference gene. Information about primers used is described in detail in **Supplementary Table S2**.

PPO activity

Protein extracts were obtained from rosette leaves homogenized at a ratio of 50 mg of fresh weight to 1000 µl of 20 mM phosphate buffer (pH 6). The homogenate was centrifuged at 16.000 × *g* and 4°C for 5 min, and the supernatant was used for protein assessment of PPO activity. PPO activity assays were performed in 20 mM phosphate buffer (pH 6) containing 10 mM L-DOPA. The reaction was measured with a SpectraMax M3 spectrophotometer (Molecular Devices, California, USA) by the change in absorbance at 475 nm and 25ºC.

Phenylpropanoid profiling

Phenylpropanoids (specifically, flavonoids) were purified and analysed as described previously[60] with minor modifications. In brief, lyophilized Arabidopsis seedling were homogenized at a ratio of 1 mg of dry weight to 33 µL of methanol:acetate:H_2O (9:1:10) extraction solvent containing 0.1 mg/mL of naringenin as an internal standard for quantification. Homogenates were incubated in the dark at 4ºC for 30 min with agitation (1000 rpm) and then centrifuged at 14,000 × *g* and 4°C for 5 min. Supernatants were recovered, filtered and analysed by high-performance liquid chromatography (HPLC) using an Agilent 1200 series HPLC equipment (Agilent Technologies, California, USA) with a XSelect CSH C18 (3.5 µm, 4.6 x 100 mm) column (Waters, Massachusetts, USA). Flavonols and anthocyanins were determined at 320 and 520 nm, respectively.

Data analysis

ANOVA followed by Newman-Keuls multiple comparison post-hoc tests were used to determine statistical significance for multiple groups. Pearson correlation coefficients (*r* values) were calculated using the means of the surface area of rosette leaves, *PPO* mRNA abundance and PPO activity. Statistical analysis was performed using GraphPad Prism 5.0a (GraphPad Software, California, USA).

References

1 Qiu, H. *et al.* Assessing the bacterial contribution to the plastid proteome. *Trends Plant. Sci.* **18**, 680-687 (2013).
2 Suzuki, K. & Miyagishima, S. Y. Eukaryotic and eubacterial contributions to the establishment of plastid proteome estimated by large-scale phylogenetic analyses. *Mol. Biol. Evol.* **27**, 581-590 (2010).
3 Price, D. C. *et al. Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science* **335**, 843-847 (2012).
4 Hernandez-Romero, D., Solano, F. & Sanchez-Amat, A. Polyphenol oxidase activity expression in *Ralstonia solanacearum*. *Appl. Environ. Microbiol.* **71**, 6808-6815 (2005).
5 Tran, L. T., Taylor, J. S. & Constabel, C. P. The polyphenol oxidase gene family in land plants: Lineage-specific duplication and expansion. *BMC Genomics* **13**, 395 (2012).
6 Wang, J. & Constabel, C. P. Polyphenol oxidase overexpression in transgenic *Populus* enhances resistance to herbivory by forest tent caterpillar (*Malacosoma disstria*). *Planta* **220**, 87-96 (2004).
7 Bhonwong, A., Stout, M. J., Attajarusit, J. & Tantasawat, P. Defensive role of tomato polyphenol oxidases against cotton bollworm (*Helicoverpa armigera*) and beet armyworm (*Spodoptera exigua*). *J. Chem. Ecol.* **35**, 28-38 (2009).
8 Thipyapong, P., Hunt, M. D. & Steffens, J. C. Antisense downregulation of polyphenol oxidase results in enhanced disease susceptibility. *Planta* **220**, 105-117 (2004).
9 Claus, H. & Decker, H. Bacterial tyrosinases. *Syst. Appl. Microbiol.* **29**, 3-14 (2006).
10 Malviya, N., Srivastava, M., Diwakar, S. K. & Mishra, S. K. Insights to sequence information of polyphenol oxidase enzyme from different source organisms. *Appl. Biochem. Biotechnol.* **165**, 397-405 (2011).
11 Marusek, C. M., Trobaugh, N. M., Flurkey, W. H. & Inlow, J. K. Comparative analysis of polyphenol oxidase from plant and fungal species. *J. Inorg. Biochem.* **100**, 108-123 (2006).
12 Jaenicke, E. & Decker, H. Functional changes in the family of type 3 copper proteins during evolution. *Chembiochem* **5**, 163-169 (2004).
13 Marchler-Bauer, A. *et al.* CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* **41**, D348-352 (2013).
14 Rubinstein, C. V., Gerrienne, P., de la Puente, G. S., Astini, R. A. & Steemans, P. Early Middle Ordovician evidence for land plants in Argentina (eastern Gondwana). *New Phytol.* **188**, 365-369 (2010).
15 Roy, S. W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet.* **7**, 211-221 (2006).
16 Stenoien, H. K. Compact genes are highly expressed in the moss *Physcomitrella patens*. *J. Evol. Biol.* **20**, 1223-1229 (2007).
17 Schonknecht, G. *et al.* Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**, 1207-1210 (2013).
18 Gaucher, E. A., Govindarajan, S. & Ganesh, O. K. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**, 704-707 (2008).
19 Kaçar, B. & Gaucher, E. A. Experimental evolution of protein-protein interaction networks. *Biochem. J.* **453**, 311-319 (2013).
20 Llorente, B. *et al.* Safety assessment of nonbrowning potatoes: opening the discussion about the relevance of substantial equivalence on next generation biotech crops. *Plant Biotechnol. J.* **9**, 136-150 (2011).
21 Emanuelsson, O., Nielsen, H. & von Heijne, G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein. Sci.* **8**, 978-984 (1999).
22 Sullivan, M. L. Beyond brown: polyphenol oxidases as enzymes of plant specialized metabolism. *Front. Plant Sci.* **5**, 783 (2014).
23 Hu, T. T. *et al*. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476-481 (2011).
24 Parfrey, L. W., Lahr, D. J., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl. Acad. Sci. USA* **108**, 13624-13629 (2011).
25 Emiliani, G., Fondi, M., Fani, R. & Gribaldo, S. A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land. *Biol. Direct* **4**, 7 (2009).
26 Sheppard, A. E. & Timmis, J. N. Instability of plastid DNA in the nuclear genome. *PLoS Genet* **5**, e1000323 (2009).

27 Stegemann, S. & Bock, R. Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus. *Plant Cell* **18**, 2869-2878 (2006).
28 Bock, R. & Timmis, J. N. Reconstructing evolution: gene transfer from plastids to the nucleus. *Bioessays* **30**, 556-566 (2008).
29 Dyall, S. D., Brown, M. T. & Johnson, P. J. Ancient invasions: from endosymbionts to organelles. *Science* **304**, 253-257 (2004).
30 Burki, F., Hirakawa, Y. & Keeling, P. J. Intragenomic spread of plastid-targeting presequences in the coccolithophore *Emiliania huxleyi*. *Mol. Biol. Evol.* **29**, 2109-2112 (2012).
31 Tonkin, C. J. *et al.* Evolution of malaria parasite plastid targeting sequences. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 4781-4785 (2008).
32 Corbin, D. R., Grebenok, R. J., Ohnmeiss, T. E., Greenplate, J. T. & Purcell, J. P. Expression and chloroplast targeting of cholesterol oxidase in transgenic tobacco plants. *Plant Physiol.* **126**, 1116-1128 (2001).
33 Nawrath, C., Poirier, Y. & Somerville, C. Targeting of the polyhydroxybutyrate biosynthetic pathway to the plastids of *Arabidopsis thaliana* results in high levels of polymer accumulation. *Proc. Natl. Acad. Sci. USA* **91**, 12760-12764 (1994).
34 Poirier, Y., Dennis, D. E., Klomparens, K. & Somerville, C. Polyhydroxybutyrate, a biodegradable thermoplastic, produced in transgenic plants. *Science* **256**, 520-523 (1992).
35 Lee, S. B. *et al.* Accumulation of trehalose within transgenic chloroplasts confers drought tolerance. *Mol. Breed.* **11**, 1-13 (2003).
36 Baudisch, B., Langner, U., Garz, I. & Klosgen, R. B. The exception proves the rule? Dual targeting of nuclear-encoded proteins into endosymbiotic organelles. *New Phytol.* **201**, 80-90 (2014).
37 Peeters, N. & Small, I. Dual targeting to mitochondria and chloroplasts. *Biochim. Biophys. Acta* **1541**, 54-63 (2001).
38 Gray, M. W. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *Proc. Natl. Acad. Sci. USA* doi: 10.1073/pnas.1421379112 (2015).
39 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).
40 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-37 (2011).
41 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).
42 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009).
43 Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307-1320 (2008).
44 Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275-282 (1992).
45 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915-10919 (1992).
46 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321 (2010).
47 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704 (2003).
48 Chevenet, F., Brun, C., Banuls, A. L., Jacq, B. & Christen, R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7**, 439 (2006).
49 Banks, J. A. *et al.* The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**, 960-963 (2011).
50 Spiegel, F. W. Contemplating the first Plantae. *Science* **335**, 809-810 (2012).
51 Geer, L. Y., Domrachev, M., Lipman, D. J. & Bryant, S. H. CDART: protein homology by domain architecture. *Genome Res.* **12**, 1619-1623 (2002).
52 Kelley, L. A. & Sternberg, M. J. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363-371 (2009).
53 Nakagawa, T. *et al.* Improved Gateway binary vectors: high-performance vectors for creation of fusion constructs in transgenic analysis of plants. *Biosci. Biotechnol. Biochem.* **71**, 2095-2100 (2007).
54 Yu, X. H. & Liu, C. J. Development of an analytical method for genome-wide functional identification of plant acyl-coenzyme A-dependent acyltransferases. *Anal. Biochem.* **358**, 146-148 (2006).

55 Karimi, M., Inze, D. & Depicker, A. GATEWAY vectors for *Agrobacterium*-mediated plant transformation. *Trends Plant Sci.* **7**, 193-195 (2002).
56 Sparkes, I. A., Runions, J., Kearns, A. & Hawes, C. Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. *Nat. Protoc.* **1**, 2019-2025 (2006).
57 Bechtold, N. & Pelletier, G. In planta *Agrobacterium*-mediated transformation of adult *Arabidopsis thaliana* plants by vacuum infiltration. *Methods Mol. Biol.* **82**, 259-266 (1998).
58 Cubo, M. T., Buendia-Claveria, A. M., Beringer, J. E. & Ruiz-Sainz, J. E. Melanin production by *Rhizobium* strains. *Appl. Environ. Microbiol.* **54**, 1812-1817 (1988).
59 Simon, P. Q-Gene: processing quantitative real-time RT-PCR data. *Bioinformatics* **19**, 1439-1440 (2003).
60 Sun, Y., Li, H. & Huang, J. R. *Arabidopsis* TT19 functions as a carrier to transport anthocyanin from the cytosol to tonoplasts. *Mol. Plant* **5**, 387-400 (2012).

Acknowledgments

We thank L. Carretero-Paulet and I. Ruiz-Trillo for critical analysis of the manuscript and all members of our lab for valuable scientific discussion. We also acknowledge the technical assistance from the CRAG core facilities staff and T. Nakagawa (Shimane University) for kindly providing the pGWB405 vector. This work was funded by grants from the European Union Marie Curie IIF program (CarotenActors), the Argentinean Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), the Spanish Dirección General de Investigación (BIO2011-23680), the Generalitat de Catalunya (XRB and 2014SGR-1434), and the Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo (IBERCAROT). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions statement

B.L. conceived the research. B.L. and M.R.-C. jointly supervised research. B.L., G.S., F.S.deS., N.D.A. and M.R.-C. designed the experiments. B.L., G.S., F.S.deS., C.M. and N.D.A performed the experiments and statistical analyses. B.L., G.S., F.S.deS., C.M., G.D.A., M.M.F., F.B.-A., N.D.A. and M.R.-C. analysed and discussed the data. B.L. drafted the manuscript and F.S.deS. and M.R.-C. contributed to the final manuscript writing. All authors read and approved the final version of the manuscript.

Additional information

Competing financial interests: The authors declare no conflict of interest.

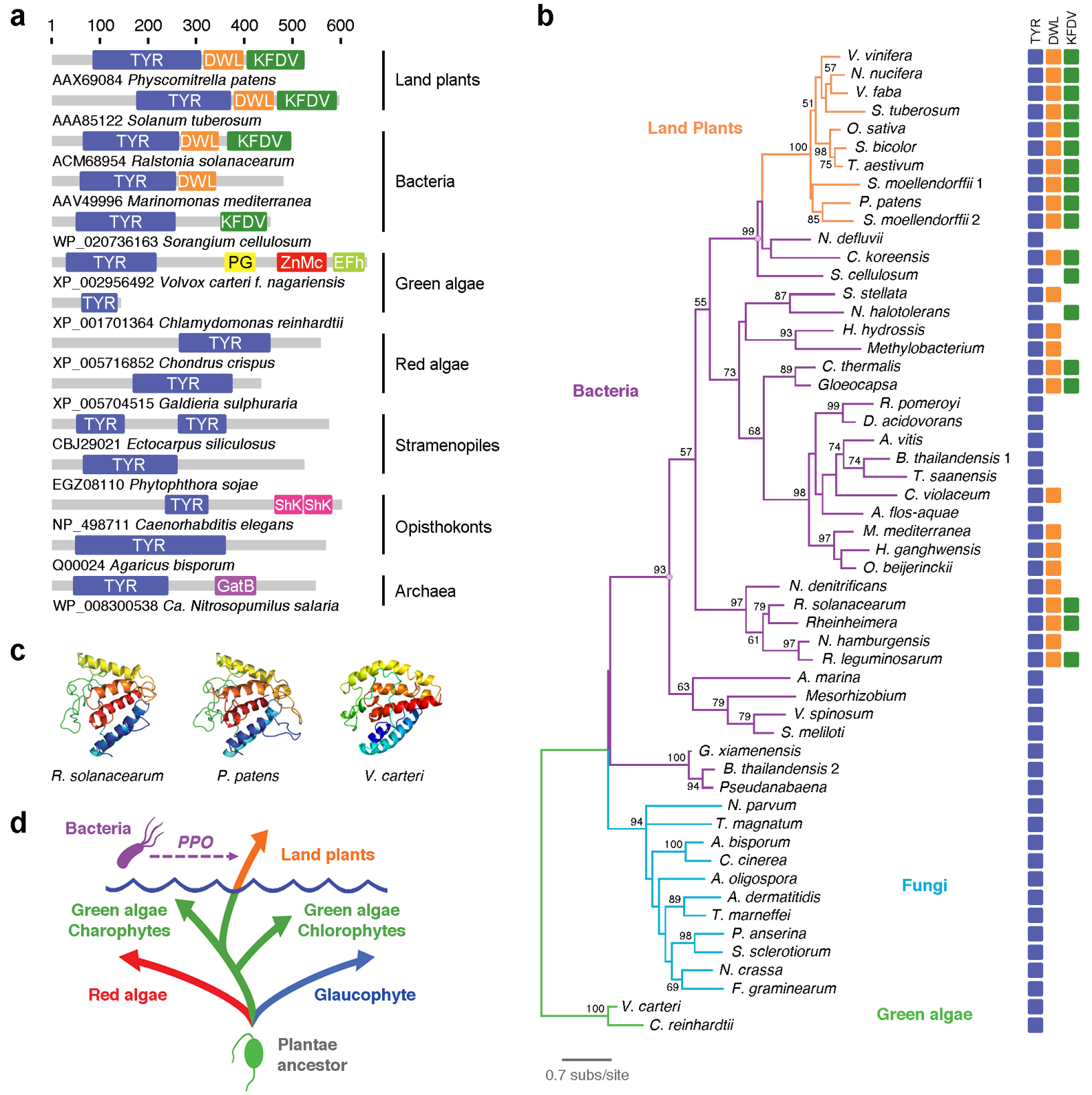
FIGURE LEGENDS

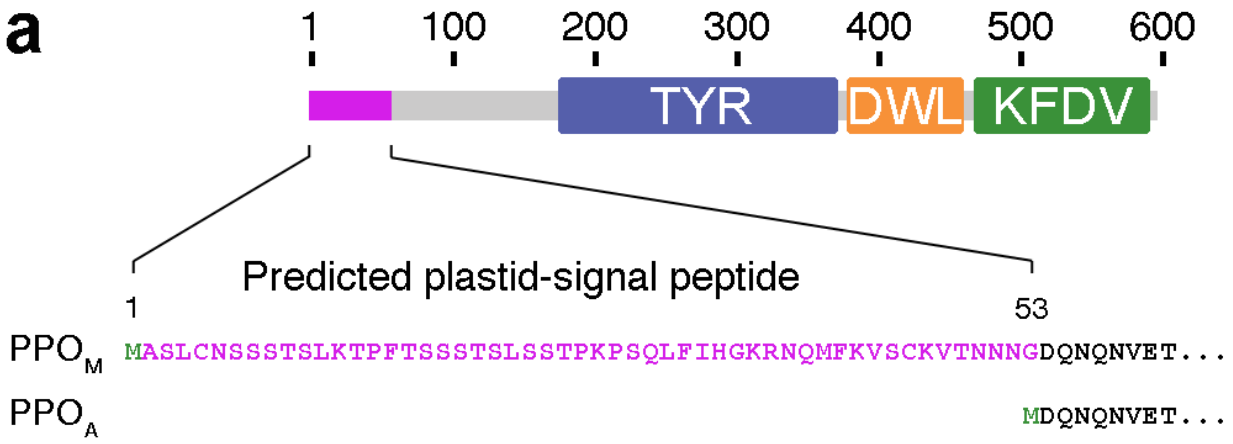
Figure 1 | Evidence for horizontal gene transfer of *PPO* from bacteria to land plants. (a) Schematic protein domain architecture alignment of different representative proteins that share the tyrosinase domain (TYR; PFAM00264). The characteristic DWL (PFAM12142) and KFDV (PFAM12143) domains of land-plant PPO enzymes are also present with the same architecture in bacterial PPO proteins but appear to be absent in green and red algae, glaucophyta and other eukaryotes. Numbers on top indicate protein (amino acids) length. PG: peptidoglycan-binding domain (PFAM01471). ZnMc: zinc-dependent metalloprotease domain (CD00203). EFh: calcium-binding domain helix-turn-helix (CD00051). ShK: ShK domain-like (PFAM 01549). GatB: GatB domain (PFAM 02637). The predicted conserved domain architecture is illustrated according to CDD results. (b) TYR domain-derived maximum likelihood inference of phylogeny showing that land plant PPO-TYR domains form a well-supported monophyletic group with several bacteria. Taxa and branches are colour coded with fungi in light blue, bacteria in purple, land plants in orange and green algae in green. Circles indicate well-supported nodes of the plant-containing bacterial branch. Bootstrap support values above 50 are shown. Domains found in the proteins of the tree are shown as boxes on the right. (c) Structures of TYR domains of *R. solanacearum*, *P. patens* and *V. carteri*. Models are independently displayed based on rainbow colouring scheme (N-terminal coloured blue and C-terminal coloured red). The analysis was performed with 100% of residues modelled at >90% confidence. (d) Scheme of the *PPO* bacteria-to-land plants horizontal gene transfer hypothesis.

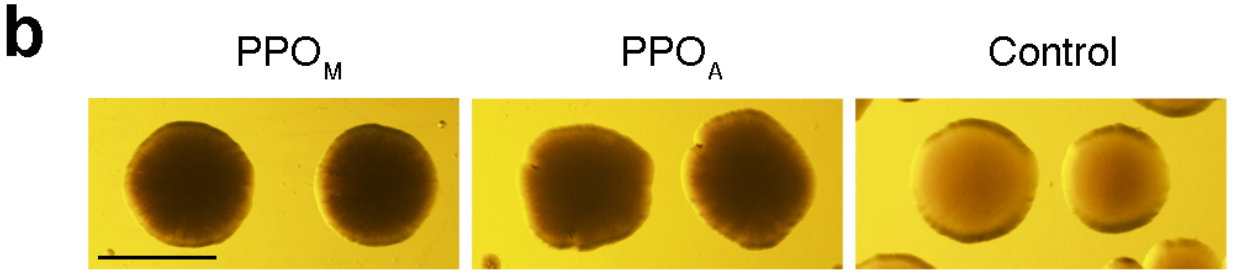
Figure 2 | Plastidial (PPO_M) and cytosolic (PPO_A) PPO models. (a) Schematic representation of the PPO enzymes used in this study. The 53 amino acids in length predicted plastid signal peptide of PPO_M (depicted in magenta) was removed from protein AAA85121 (597 amino acids in length) to generate PPO_A. Numbers on top indicate protein (amino acids) length. (b) Melanin production in *E. coli* cells expressing PPO_M and PPO_A compared with cells transformed with control plasmid. Images correspond to 3-day-old *E. coli* colonies grown on LB supplemented with CuSO4, IPTG and chlorogenic acid. Scale bar represents 200 μm. (c) Subcellular localization of PPO_M (plastids) and PPO_A (cytosol) fused to GFP in *N. benthamiana* leaves visualized by confocal microscopy. Scale bars represent 10 μm.

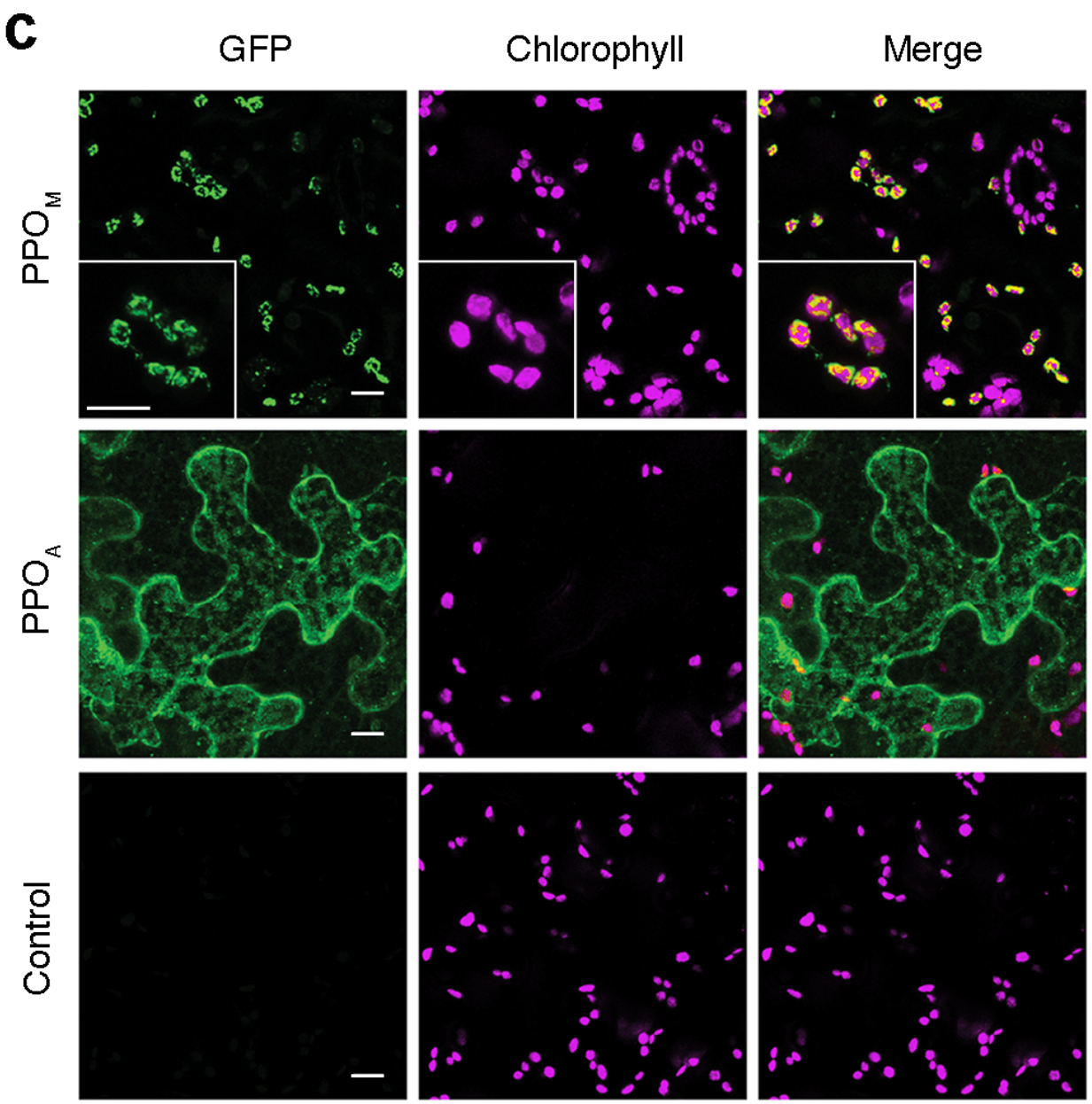
Figure 3 | Arabidopsis plants expressing PPO_A have reduced growth rate and delayed bolting time. (a) RT-qPCR analysis of *PPO* mRNA abundance in PPO_M and PPO_A lines. Data are represented relative to PPO_M line 7 and correspond to mean ± SEM derived from three technical repeats of three biological replicates. Data correspond to 4-week-old plants. (b) PPO activity for PPO_M and PPO_A lines. Data are represented relative to PPO_M line 7 and correspond to mean ± SEM derived from three technical repeats of three biological replicates. Data correspond to 4-week-old plants. (c) Phenotype of PPO_M and PPO_A lines. Data correspond to 4-week-old plants. (d) Surface area of rosette leaves for PPO_M and PPO_A lines. Data correspond to mean ± SEM derived from 10 biological replicates. Asterisks represent statistically significant differences (***$P = 0.001$) according to the ANOVA followed by Newman-Keuls post-hoc test. Data correspond to 4-week-old plants. (e) Bolting time in PPO_M and PPO_A lines. Data correspond to mean ± SEM derived from 10 biological replicates. Asterisks represent statistically significant differences (*$P = 0.05$, ***$P = 0.001$) according to the ANOVA followed by Newman-Keuls post-hoc test.

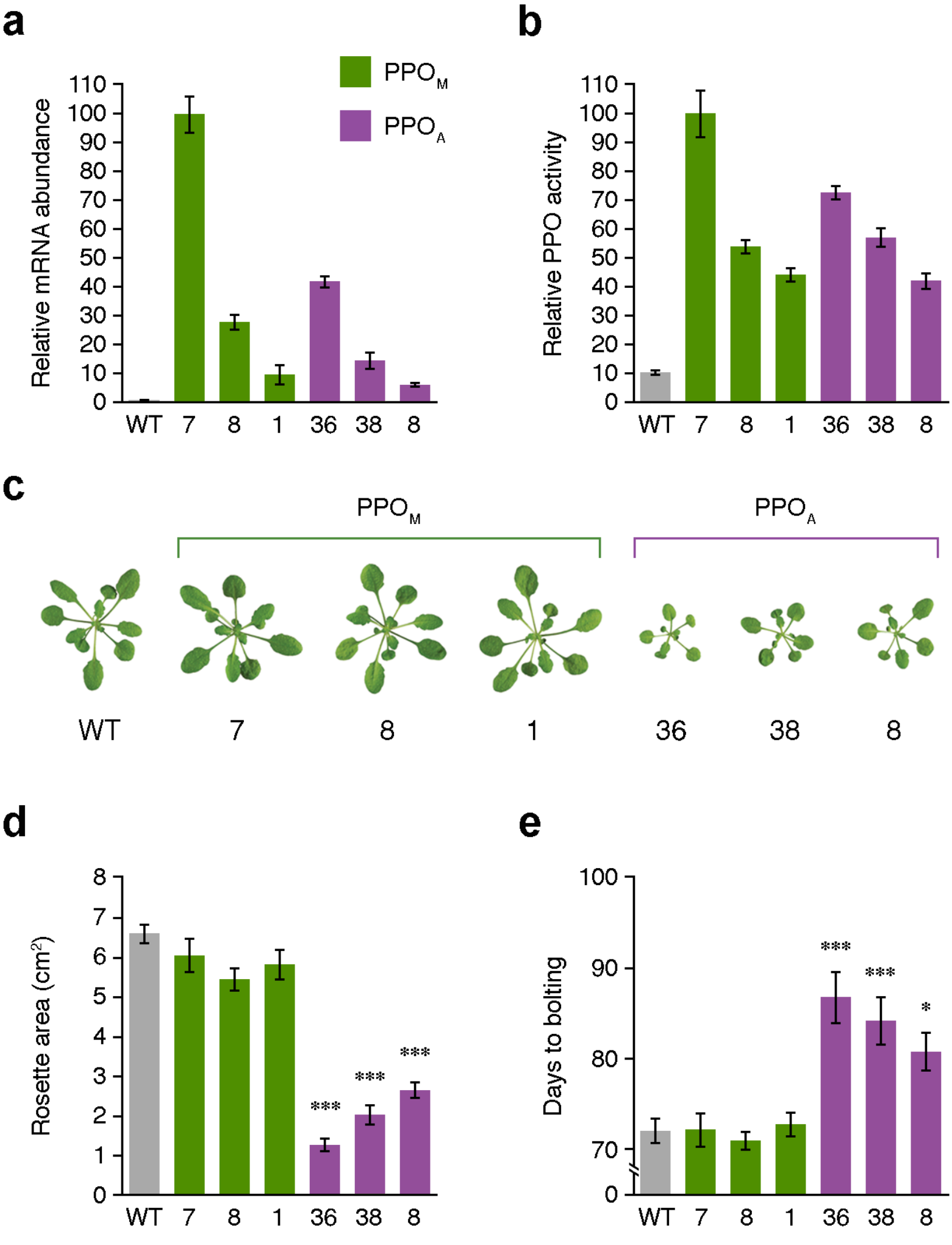
Figure 4 | Qualitative and quantitative profiles of phenylpropanoid metabolites in transgenic and untransformed Arabidopsis seedlings. Transgenic PPO_M and PPO_A lines together with untransformed (WT) controls were grown on sucrose-supplemented medium and used for phenylpropanoid profiling as described in Materials and methods. (a) Representative HPLC chromatograms of flavonols. (b) Representative HPLC chromatograms of anthocyanins. (c) Total flavonol and anthocyanin levels represented relative to WT samples. Data correspond to mean ± SEM derived from three technical repeats of four biological replicates. Asterisks mark statistically significant differences (***$P = 0.001$) according to the ANOVA followed by Newman-Keuls post-hoc test. (d) Phenotype of seedlings grown for 10 days in the presence of norflurazon. Note the stronger anthocyanin (purple) pigmentation in PPO_A seedlings. Scale bar represents 0.5 cm.

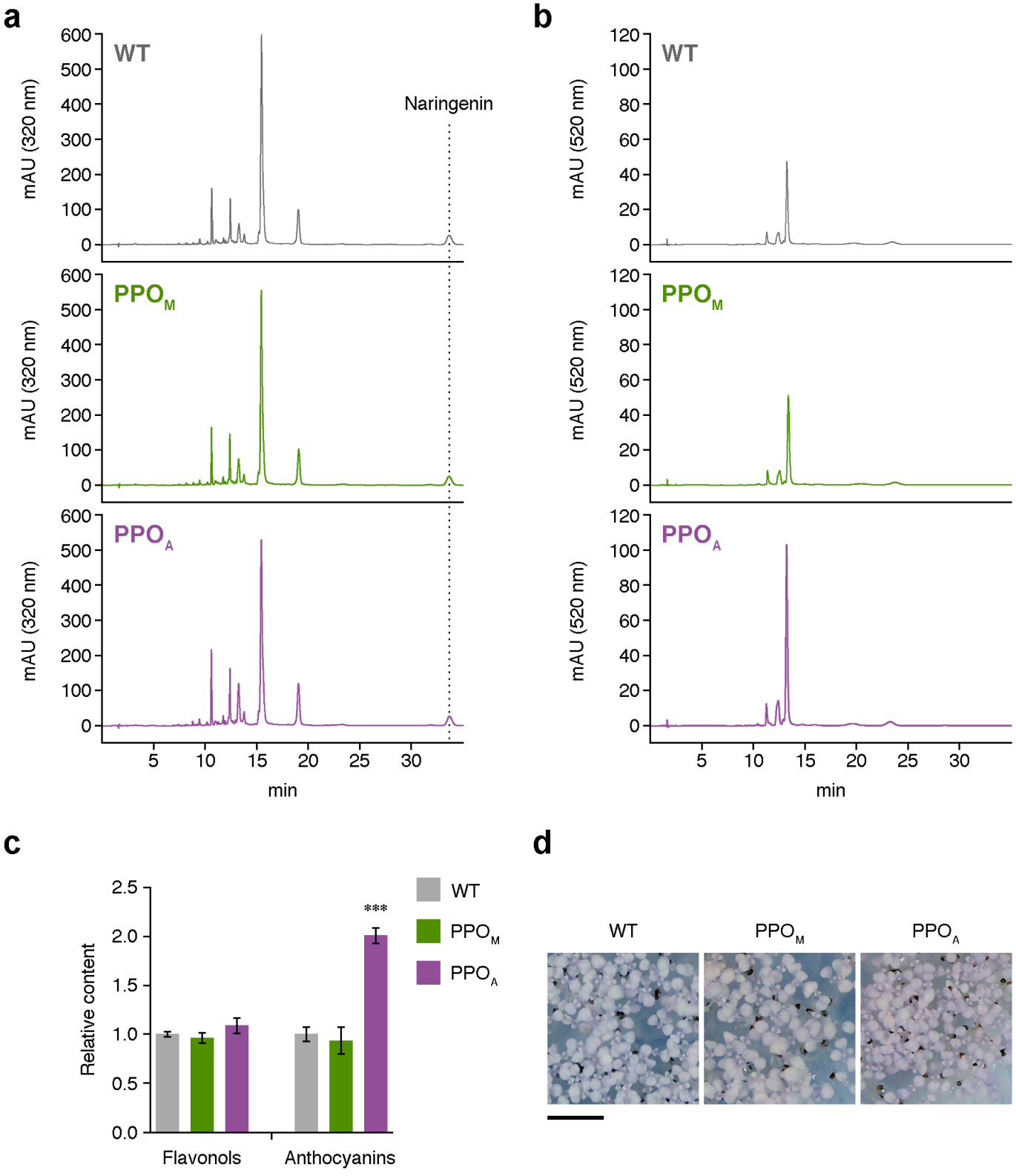

a
1
100
200
300
400
500
600
TYR
DWL
KFDV
AAX69084 Physcomitrella patens
AAA85122 Solanum tuberosum
ACM68954 Ralstonia solanacearum
AAV49996 Marinomonas mediterranea
WP_020736163 Sorangium cellulosum
PG
ZnMc
EFh
XP_002956492 Volvox carteri f. nagariensis
XP_001701364 Chlamydomonas reinhardtii
XP_005716852 Chondrus crispus
XP_005704515 Galdieria sulphuraria
CBJ29021 Ectocarpus siliculosus
EGZ08110 Phytophthora sojae
ShK
NP_498711 Caenorhabditis elegans
Q00024 Agaricus bisporum
GatB
WP_008300538 Ca. Nitrosopumilus salaria
Land plants
Bacteria
Green algae
Red algae
Stramenopiles
Opisthokonts
Archaea
b
TYR
DWL
KFDV
Land Plants
V. vinifera
N. nucifera
V. faba
S. tuberosum
O. sativa
S. bicolor
T. aestivum
S. moellendorffii 1
P. patens
S. moellendorffii 2
Bacteria
N. defluvii
C. koreensis
S. cellulosum
S. stellata
N. halotolerans
H. hydrossis
Methylobacterium
C. thermalis
Gloeocapsa
R. pomeroyi
D. acidovorans
A. vitis
B. thailandensis 1
T. saanensis
C. violaceum
A. flos-aquae
M. mediterranea
H. ganghwensis
O. beijerinckii
N. denitrificans
R. solanacearum
Rheinheimera
N. hamburgensis
R. leguminosarum
A. marina
Mesorhizobium
V. spinosum
S. meliloti
G. xiamenensis
B. thailandensis 2
Pseudanabaena
N. parvum
T. magnatum
A. bisporum
C. cinerea
A. oligospora
A. dermatitidis
T. marneffei
P. anserina
S. sclerotiorum
N. crassa
F. graminearum
Fungi
V. carteri
C. reinhardtii
Green algae
0.7 subs/site
c
R. solanacearum
P. patens
V. carteri
d
Bacteria
PPO
Land plants
Green algae
Charophytes
Green algae
Chlorophytes
Red algae
Glaucophyte
Plantae
ancestor


a
1
100
200
300
400
500
600
TYR
DWL
KFDV
Predicted plastid-signal peptide
1
53
PPO_M MASLCNSSSTSLKTPFTSSSTSLSSTPKPSQLFIHGKRNQMFKVSCKVTNNNGDQNQNVET...
PPO_A MDQNQNVET...


b
PPO_M
PPO_A
Control


c
GFP
Chlorophyll
Merge
PPO_M
PPO_A
Control


a
PPO_M
PPO_A
Relative mRNA abundance
110
100
90
80
70
60
50
40
30
20
10
0
WT 7 8 1 36 38 8
b
Relative PPO activity
110
100
90
80
70
60
50
40
30
20
10
0
WT 7 8 1 36 38 8
c
PPO_M
PPO_A
WT 7 8 1 36 38 8
d
Rosette area (cm²)
8
7
6
5
4
3
2
1
0
***
***
***
WT 7 8 1 36 38 8
e
Days to bolting
100
90
80
70
0
***
***
*
WT 7 8 1 36 38 8


a
WT
Naringenin
PPO_M
PPO_A
mAU (320 nm)
min
b
WT
PPO_M
PPO_A
mAU (520 nm)
min
c
Relative content
WT
PPO_M
PPO_A
***
Flavonols
Anthocyanins
d
WT
PPO_M
PPO_A